DOCUMENT RESUME

ED 051 752 HE 002 226

AUTHOR	Spencer, Richard E.; Aleamoni, Lawrence M.
TITLE	The Illinois Course Evaluation Questionnaire: Description of its Development and a Report of Some of its Results. A History of the Illinois Course Evaluation Questionnaire.
INSTITUTION	Illinois Univ., Urbana. Office of Instructional Resources.
REPORT NO	RR-292; RR-306
PUB DATE	Sep 69
NOTE	36p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Course Evaluation, *Evaluation, *Higher Education, *Questionnaires, Student Opinion
IDENTIFIERS	*Illinois Course Evaluation Questionnaire

ABSTRACT

This document consists of two reports. The first report discusses the purpose in developing the Illinois Course Evaluation Questionnaire (CEQ), which was to devise an instrument that could elicit student opinions about a standardized set of statements relative to certain standardized aspects of an instructional program, and to develop norms which would enable the instructor to adequately compare his results with the results of other instructors. It also reviews the method of the questionnaire development: the criteria applied, the experience of other universities, and the item selection. The questionnaire's subscore development, its reliability and the scoring procedure are explained in the appendix, where the development of the normative data, the system for reporting and developing the questionnaire results, and report of some studies which have used the CEQ are also discussed. The second report expands on the development aspects of the questionnaire. (AF)

ED051752

research report # 292

research report

The Illinois Course Evaluation Questionnaire: Description of Its Development and a Report of Some of Its Results

by

Richard E. Spencer and Lawrence M. Aleamoni

Measurement and Research Division
Office of Instructional Resources
University of Illinois
507 East Daniel Street, Champaign

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

measurement and research division
office of instructional resources
university of illinois
urbana, illinois

HE 002 226

ED051752

The Illinois Course Evaluation Questionnaire: A Description of Its Development and a Report of Some of Its Results

Richard E. Spencer and Lawrence M. Aleamoni
University of Illinois

If one assumes that the purpose of education is to change students behavior as a result of some definite course of instruction, then an objective of educational research should be to determine what procedures or techniques best produce the desired behavioral changes. If the desired behavioral changes can be identified and defined then the educational researcher can develop instruments to measure them.

Let us also assume that if one does in fact change student behavior, in the specified direction, as a result of a course of instruction, then that course has been effective. If that course has been effective then there could be a large number of elements in that course contributing to its effectiveness, such as the instructor, textbook, homework, course content, method of instruction, student interest, student attention, general student attitude towards the course, etc.

Assuming that all of the elements enumerated above can affect, directly or indirectly, student behavior in a course, and assuming that the students are the only ones who are constantly exposed to those elements, then they appear to be the most logical evaluators of the quality and effectiveness of the course elements. In addition, student opinions should indicate areas of rapport, degrees of communication, or the existence of problems and thereby help instructors as well as educational researchers describe and define the learning environment more concretely and objectively than they could through other types of measurements.

There are various ways of sampling student opinion. Some useful information can be derived simply be determining the number of students who agree or disagree with certain statements about the course. Or, sometimes, it proves useful to ask students to write short essays about the course in order to obtain information about their experiences under specific instructional situations. Such individualized procedures do not, however, provide an opportunity to compare the results of one course with results of another. Measurement is more useful when comparative results are available. More adequate interpretation may occur when: (a) the data has been collected in a standardized fashion with appropriate attention given to sampling, reliability, and validity, and (b) many instructors and instiuctional programs have been measured with the same instrument so that comparisons can be made.

Therefore, the primary purpose in developing the Illinois Course Evaluation Questionnaire (CEQ) was to devise an instrument which could elicit student opinions about a standardized set of statements relative to certain standardized aspects of an instructional program, and to develop norms which would enable an instructor to adequately compare his results with the results of other instructors.

METHOD OF QUESTIONNAIRE DEVELOPMENT

Initially, questionnaires developed at other universities (Anderson, 1954; Bradley, 1950; Coffman, 1954; Cosgrove, 1959) were considered, however, they all seemed to suffer the same disadvantages, such as inadequate sampling, lack of validity data, and lack of normative data. For example, certain questionnaires were designed to collect attitudes on the instructor

only, evidently under the assumption that other variables inherent in the learning situation would not significantly effect the potential role of the instructor in affecting learning. On the other hand, the present questionnaire was developed under the assumption that the instructor is only capable of influencing the learning situation to the degree that he is not restricted by elements outside of his control. Some of these elements would include scheduling, grouping, course content, curriculum or college requirements, and previous student opinions. It seemed necessary, therefore, to develop an instrument which would tap the attitudes of students differentially; that is, to obtain results on those elements within the learning situation which relate to learning as well as to teaching. It seemed possible that an instructor might teach certain content excellently, but opinions about his teaching effectiveness could be prejudiced by the attitudes toward the content of the course per se, therefore, the measuring instrument should test these elements separately.

A review of the various procedures and forms for collecting student opinions (Anikeeff, 1953; McKeachie & Solomon, 1958; Patton & Meyer, 1955; Remmers, 1960; Weaver, 1960) opinions indicated that certain limitations should be imposed before selecting or constructing an appropriate instrument. The following criteria appeared relevant and were applied:

1. Administration: The questionnaire should be administered by the instructor himself, during the regular class or examination time, so that proctors and administrators would not be necessary.
2. Time: It should be short enough to be acceptable to faculty in regular classes, but long enough to insure reliability and an adequate measure of a wide sample of attitudes.

3. Content: It should measure those opinions and attitudes which are developed or exist about the total instructional program rather than a single element therein. It should also not measure invalid elements such as dress, room temperature, etc.
4. Scoring: It should be objective, and preferably machine scorable so that the results could be returned promptly and scoring could be standardized and reliable.
5. Reliability: If one wants to insure that scores on the instrument are a true representation of the students' opinions, those scores must be reproduceable upon subsequent testing of the same students rating the same instructor and course.
6. Interpretation: It should yield scores which differentiate among instructional programs, and which can be interpreted by instructors in such a manner that their instructional effectiveness can be improved. It should assist in the diagnosis of the strengths and weaknesses of the instructional program.
7. Realism: The attitudes measured must include those critical elements which comprise the opinions the student has and wishes to express; but the measuring instrument must be capable of eliciting "real" feelings, and not careless or merely socially acceptable or expected responses.

Criteria for effective instruction were culled from the extensive literature (Fults, 1952; Gouds, 1960; Justman & Mais, 1956) on the subject and then items were constructed and assigned to the various criteria on the basis of their face validity. Approximately 150 separate items were developed in this manner. Additional items were obtained through the work of a

faculty committee at the Pennsylvania State University investigating effective instruction. A student committee also at Pennsylvania State was asked to submit items. As a result, a pool of over 1,000 items was obtained and then administered to 1,200 undergraduate students and the Pennsylvania State University in Fall 1962. The response scale for these items consisted of five points (strongly agree, agree, uncertain, disagree, strongly disagree). In the resulting analysis many items were deleted because they appeared to be measuring much the same thing as other items, and some were dropped because they did not differentiate among instructors, thereby leaving a total of 450 items. The 450 items were then administered to another sample and reanalyzed, yielding a third reduced form containing 100 items. This form was administered to 1,319 undergraduate students in accounting, anthropology, army ROTC, history, mathematics, music, psychology, and zoology courses at the Pennsylvania State University.

The number of items continued to be reduced by further reanalyses involving the use of factor analysis until 28 relatively distinct items were obtained. Results of the above analyses indicated that a large number of students responses were falling at the neutral position on the response scale and that careless and invalid responding was the rule on a significant proportion of the questionnaires. Therefore, a forced choice answering technique was adopted to handle the scale problem by eliminating the neutral position. In addition a response set score was developed to handle the careless responses by constructing 22 negatively stated items that expressed roughly the same concepts as 22 (out of the 28) corresponding positively stated items. This, then, resulted in a final version of the questionnaire (CEQ) which contained only 50 items.

Appendix A

Sub-Score Development

Factor analysis (Thurstone, 1947) was used on the initial and all other versions of the CEQ and resulted in the same six sub-scores (or factors) being defined as found in Table 1.

TABLE 1

Factors Obtained from the Fianl 50 Item Questionnaire

	Factor (Sub-Score)	Number of Items In the Factor	% Variance
I	General Course Attitude	8	30
II	Method of Instruction	8	6
III	Course Content	8	5
IV	Interest and Attention	8	4
V	Instructors	8	3
VI	Other	10	3

The percent of variance in student scores which is generally accounted for by each of the factors is shown in Table 1 for the initial 1,200 sample and has remained relatively the same for all subsequent factor analysis of the CEQ. The number of questionnaire items (in the final version) composing each factor is also indicated.

The sub-score correlations (VII represents the Total Score) also based upon the initial administration of the questionnaire to the 1,200 students are presented in Table 2 and clearly indicate that there is a high relationship between the sub-scores. However, since the correlations between the scores of any two sub-scores are generally lower than the reliability

of either of the sub-scores, it may be concluded that each of the sub-scores is measuring, in part, something which is unique. These correlations also remained stable when calculated for subsequent versions of the questionnaire.

TABLE 2

Correlations Among Sub-Scores

	I	II	III	IV	V	VI	VII
I	1.00						
II	.67	1.00					
III	.72	.65	1.00				
IV	.77	.69	.64	1.00			
V	.47	.55	.46	.52	1.00		
VI	.61	.68	.73	.60	.49	1.00	
VII	.86	.87	.83	.87	.69	.82	1.00

The fact that "General Course Attitude" accounts for the largest proportion of variance and the sub-scores are highly related indicates that there is probably some general factor underlying the responses.

Reliability

The split-half method (Guilford, 1956) of calculating reliability was used on the negative vs. positive items of a sample of 297 CEQ's, yielding a correlation of .849, which when corrected for length by the Spearman-Brown formula (Guilford, 1956) resulted in a correlation of .92. A second

split-half reliability was computed with half the negative and half the positive items in each group; thus 25 items in each half. The result was an obtained correlation of .865, which when corrected for length equalled .93.

In addition, the Kuder-Richardson (1937) reliability, formula 21 (K-R 21), was computed for 16 different courses, which resulted in an average K-R 21 of .931 and a standard deviation of .02. Since the K-R 21 has the underlying assumptions of: (a) a single common factor, (b) all inter-item correlations being equal, (c) scoring formula being the number of correct responses, and (d) the item difficulties being equal, the most positive response to each questionnaire item was assumed to be a "right" answer and all others wrong which, therefore, would provide an underestimate of the reliability of the questionnaire.

The responses of students in seven additional courses were used to determine the reliabilities of the sub-scores. The reliabilities, as computed by K-R 21, are presented in Table 3.

TABLE 3

K-R 21 Reliabilities of Sub-Scores for Seven Selected Courses

	Sub-Score	Courses A	B	C	D	E	F	G
I	General Course Attitude	.845	.845	.737	.782	.828	.790	.708
II	Method of Instruction	.924	.864	.743	.777	.836	.837	.797
III	Course Content	.556	.657	.395	.539	.672	.581	.508
IV	Interest and Attention	.894	.846	.762	.779	.827	.815	.709
V	Instructors	.721	.768	.645	.724	.771	.725	.731
VI	Other	.569	.700	.500	.629	.737	.680	.521
	N	460	146	340	189	296	571	94

The reliability coefficients in Table 3 indicate that the items defining the "Content" and "Other" sub-scores are the least reliable. Since a few of the content items correlated with the general items, this would explain the lower internal consistency of the items. The "Other" items were chosen because of their specific, and not necessarily related, content and, therefore, would not be expected to be highly inter-correlated. The reliabilities of the other four sub-scores would generally be considered acceptable.

SCORING PROCEDURE

The CEQ is scored using a weighted point system, under the assumption that students who indicate strong responses to questionaire items should be differentiated from those whose responses tend to be more moderate. All responses to the CEQ items are based on a common scale, from strongly agree (SA), through agree (A), to disagree (D), and strongly disagree (SD). There is no neutral position. Since there are CEQ items which express positive or negative attitudes toward the instructional program, these items have to be scored differently inasmuch as agreeing with a positive item would indicate a positive attitude toward the course, while agreeing with a negative item would indicate a negative attitude.

A response-set score was developed by matching items expressing roughly the same concept and are presented in Table 4. There was one positive and one negative item in each pair. The scoring of these items in matched pairs is useful in identifying the "careless" student responses. Such results can be identified in the scores for any instructor. The response-set score is also helpful in explaining score unreliability resulting from the failure of students to know their true opinions or to express them honestly.

TABLE 4

Matched Positive and Negative Items in each of the Sub-Scores

	Sub-Score	Matched Pairs with Negative Items Underlined	Unmatched Items
I	General Course Attitude	3-2,20-34,25-11,49-29	
II	Method of Instruction	6-37,27-48, 36-8,50-1	
III	Course Content	30-28,40-44,13-39,19-26	
IV	Interest and Attention	7-24,9-14,22-46,35-45	
V	Instruction	5-31,12-23,18-10,47-15	
VI	Other	21-41,42-33	4,16,17,32,38,43

The CEQ's were studied for the total number of responses filled in by the students, versus the number of items omitted. Since the scoring system uses weights, an omit would affect the total score obtained. Some individual students have been found to leave all items blank. One section was found to have left 30% of the items blank.

It was therefore, decided to score each item according to the number of students who answered it. Thus, the mean score for each item may come from slightly different size-samples of students. These numbers are reported in the Summary Report to Instructors. It should be understood, however, that the reliability of scores on the questionnaires are related to the size of the group tested, therefore, small group results should be considered highly tentative.

In the scoring procedure, the average item response is computed for each item for a given class. The instructors' item means are then compared to the total results across all sections tested in the standardization population and decile norms are printed for each item mean. A total score and a set of sub-scores are also computed and presented in the report to the instructor. The total score is the mean response over all questionnaire items. The sub-scores represent definite areas in an instructional program that can be considered relatively independent from each other. For example, the content of a course may be rated by the students as good while the method of instruction may be considered poor. Thus, a total score might disguise the various parts of an instructional program that may be viewed differentially by the students.

NORMATIVE DATA

In the initial development of the norms, 406 sections (7,083 students in all) at the University of Illinois and The Pennsylvania State University were given the present version of the questionnaire during the Fall Semester 1964-1965. In June 1965, 364 sections were given the questionnaire. To date, approximately 250 courses with a total of approximately 800 sections and a total of over 100,000 students represent the University of Illinois normative population. The Pennsylvania State University data are not included in the norm population.

Normative data for each item, expressed in deciles, is based upon the responses of the total normative population whereas the normative data for the subscores is also reported by department; level, rank of instructor, etc. The normative data is continuously being up-dated with each new semester's results.

In addition, administrations of the questionnaire to representative samples (sometimes exhaustive) at Temple University, Eureka Jr. College, Bowling Green University, University of Alabama, Chicago Circle Campus of the University of Illinois, and University of Oregon indicate that the normative data is relatively invariant from institution to institution.

REPORTING AND INTERPRETING THE QUESTIONNAIRE RESULTS

A system has been developed to automate the processing of the Illinois Course Evaluation Questionnaire. When a request is recieved for the use of the questionnaire, the user is provided with a copy of the form which is printed on a Digitek Answer sheet so that the students can respond by marking directly on the questionnaire sheet. (See Appendix A). It normally takes about 10 minutes to complete the CEQ.

Punched cards are then produced from the answer sheets and submitted with a computer program to produce results for a particular class. The results include:

A. a print-out which indicates average sub-test and total scores, and the norm decile, and

B. a print-out which includes specific item responses, their means and the norm decile.

Two copies of the results are returned only to the instructor, but pooled results for entire courses consisting of many sections may also be obtained with instructor identification eliminated. As the number of measures on each course is increased, it becomes possible to obtain a relatively stable indication of the difference between courses. This aids in the interpretation of the actual differences between an obtained section score for a particular instructor and the average scores for all the sections represented in that course.

REPORT OF SOME STUDIES

Although many studies have been conducted using the CEQ only four will be mentioned below.

A question of immediate interest was that of the relationship between the various sub-scores and the variables consisting of sex, term, curriculum, and final grade. The questionnaire was administered to two courses at the Pennsylvania State University and the student information was obtained for approximately 300 students in each course (Spencer & Dick, 1965). The correlations between the student information and the sub-scores were computed for both courses.

The results indicated that the responses to the questionnaire had little or no relationship to the student's sex, term, or curriculum. However, course grade and scores on the questionnaire did correlate significantly (even though the magnitude of the correlations was small) with all the sub-scores except the instructor rating. These results, plus previous research by other investigators (Remmers, 1960; Weaver, 1960), indicated that course grades do correlate with course evaluations, but the correlation seldom exceeds .30.

In another study by Spencer and Dick (1965), however, the Course Evaluation Questionnaire was administered to 12 sections in Speech 101, at the Pennsylvania State University from which 160 student responses were obtained. The questionnaire was administered during the semester (2 weeks after midterm). Two sub-scores on the questionnaire were used in a comparison with four validating criteria of "Success in Speech." The following correlations were obtained:

1. Student Attitude toward instructor, and grades obtained on 6 class speeches, r = .85.

2. Student attitude toward instructor, and a Test of Principles of Disposition Form A, r = .91.

3. Student attitude toward instructor, and a Test of Principles of Disposition Form B, r = .90.

4. Student attitude toward instructor, and the Goyer Organization of Ideas Test, r = .36.

The sub-score of "Method of Instruction" correlated slightly with Form A and B Test of Principles of Disposition.

It can be seen, then, that in some courses, student opinion about the course is highly related to success in the course.

A study involving the use of anonymous and identified student responses to the CEQ by Spencer (1965) indicated that students do answer differently when they are asked to identify themselves.

Stallings and Spencer (1967) compared the judgements of 10 raters viewing nine instructors teaching Accountancy 101 via video-tape clips to the instructors CEQ total score ratings. They found a significant correlation (p = .70) between the CEQ total score ranks and the average rating ranks for the nine instructors.

DISCUSSION AND SUMMARY

The measurement of the effectiveness of instruction is a complex problem. It may be approached in various ways. The CEQ was designed to collect evidence of only one kind--student opinion, which appeared to be the most relevant kind.

This questionnaire has definite advantages over those similarly oriented in that large representative samples have been obtained upon which norms have been established to provide course, section, department, etc. comparisons. The Inter-university comparisons established the generality

of the norms and the matched positive and negative items provide an excellent lie-score test. The ease of administering, scoring, and interpreting the results also add to its attractiveness.

It would seem, on the basis of the face validity of the CEQ and its high reliability, that extremely low scores on a particular sub-score should indicate "felt" problem areas in an instructor's teaching procedure. Whereas, stable high scores should point to an effective instructional program as viewed by students. All available validating evidence, to date, indicates that the CEQ does indeed identify courses that are considered to be very good or very bad.

The results of the factor analysis of items and the sub-score interrelationships indicated that no one element, related to a course, disproportionately influenced the students' evaluation of the course. It appears that there is a "general course attitude" cultivated by the student as he is exposed to previous student's comments, the instructor, the textbook, the course, etc. and this is the framework from which he responds when answering the CEQ items.

Variants of this questionnaire have been constructed for use with high school students, student teachers and anecdotal information collected from college students. Studies are being conducted on these alternate forms to see if anything unique is being obtained about course evaluation.

References

Anderson, C. L. The student looks at his learning. Improving College and University Teaching, 1954 (November), 2, 65-66.

Anikeeff, A. M. Factor affecting student evaluation of college faculty members. Journal of Applied Psychology, 1953, 37, 458-460.

Bradley, Glandyce H. What do college students like and dislike about college teachers and their teaching? Educational Administration and Supervision, 1950 (February), 36, 113-120.

Coffman, W. E. Determining students' concepts of effective teaching from their ratings of instructors. Journal of Educational Psychology, 1954, 45, 277-286.

Cosgrove, D. J. Diagnostic rating of teacher performance. Journal of Educational Psychology, 1959, 50, 200-204.

Fults, Anna C. Evaluating college teaching. Journal of Home Economics, 1952 (January), 44, 21-22.

Goods, D. The centrality of evaluation. Improving College and University Teaching, 1960, 8, 16-18.

Guilford, J. P. Fundamental Statistics in Psychology and Education. (3rd ed.) New York: McGraw-Hill, 1956.

Justman, J. & Mais, W. H. College teaching: its practice and potential. New York: Harper & Bros. 1956.

Kuder, G. F. & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

McKeachie, W. J. & Solomon, D. Student ratings of instructors. Journal of Educational Research, 1958, 51, 379-382.

Patton, R. M. & Meyer, P. R. A forced choice rating form for college teachers. Journal of Educational Psychology, 1955, 46, 499-503.

Remmers, H. H. Manual of instructions for the Purdue rating scale for instruction. West Lafayette, Indiana: University Book Store, 360 State Street, 1960.

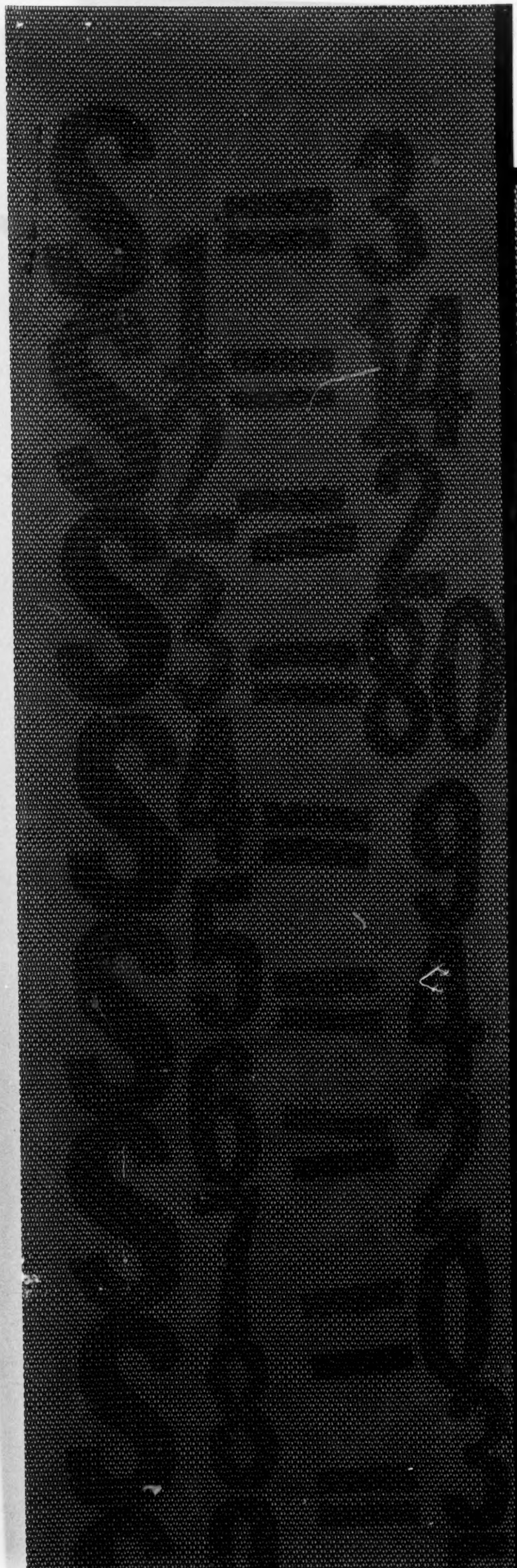
Spencer, R. E. Anonymous vs. identified student responses. Research Study No. 202, Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1965.

Spencer, R. E. & Dick, W. Course evaluation questionnaire: manual of interpretation. Research Report No. 200, Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1965.

Stallings, W. M. & Spencer, R. E. Ratings of instructors in Accountancy 101 from video-tape clips. Research Report No. 265, Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1967.

Thurstone, L. L. Multiple-factor analyses: a development and expansion of the vectors of mind. Chicago: University of Chicago Press, 1947.

Weaver, C. H. Instructor rating by college students. Journal of Educational Psychology, 1960, 51, 21-25.



306

research report #

research report

A History of the Development of the Illinois Course Evaluation Questionnaire

by

Richard E. Spencer

Measurement and Research Division
Office of Instructional Resources
University of Illinois
507 East Daniel Street, Champaign

September 1969

measurement and research division
office of instructional resources
university of illinois
urbana, illinois

20

A History of the Development of the

Illinois Course Evaluation Questionnaire

The Illinois Course Evaluation Questionnaire (CEQ)

College level. Copyright 1965, Richard E. Spencer, The University of Illinois, Urbana, Illinois. DIGITEK answer sheets serve as test booklets, 50 item plus identification information, manual (10 pages), 15-20 minutes, scoring service. Available through the Measurement and Research Division, Office of Instructional Resources, University of Illinois. Optional answer positions available for up to 50 additional or locally constructed questions. Identification data on the form includes:

Student ID number (9 digits)

Course Code (5 digits)

Expected grade in this course (A to E)

Is this course required or elective

Sex of the student

College of the student

Date (Month, Day, Year)

Student Status (Freshman, Sophomore, etc)

A short form of the Illinois Course Evaluation Questionnaire is also available, containing only 25 items (all the positive items), as well as the appropriate identification data.

Norms are available for more than 100,000 students 2,000 course sections, and 400 different courses. Differential norms available by all university classes, rank of instructor, level of course (4 levels), college, department, and certain content areas. Norms include data from:

The University of Illinois - Urbana, Illinois

Bowling Green State University - Bowling Green, Ohio

Temple University - Philadelphia, Pennsylvania

Maritime Academy - Bronx, New York

Shippensburg State University - Shippensburg, Pennsylvania

Eureka College - Eureka, Illinois

Lake Land College - Mattoon, Illinois

University of Iowa - Iowa City, Iowa

University of Oregon, Eugene, Oregon

University of Alabama - University, Alabama

University of South Alabama - Mobile, Alabama

Freed-Hardeman College - Henderson, Tennessee

Illinois State University - Normal, Illinois

University of Michigan, Dearborn - Dearborn, Michigan

Scoring service (answer sheets, manual, I.B.M. card preparation, scoring, and reporting) available from M&R, University of Illinois.

Reliabilities (instrument and rater) are computed for each score and sub-score for each class section processed, and printed on the computer output. Two copies are returned to the user.

History of the Questionnaire:

The concept of teacher or instructional rating scales is aptly stated by Dale Wolfe.

> "... the basic problem cannot be solved unless the status of teaching is enhanced in the eyes of present and prospective faculty members and the supporters of higher education."
>
> "... if great teaching is to be rewarded, the great teachers must be identified. And here there is a problem for those who contend that the quality of teaching is unmeasurable."
>
> "As a start, it should be possible on any campus to collect independent ratings, preferably on first-hand evidence rather than on hearsay. If it turns out that there is reasonably high consistency in the judgments, good; the point has been made that the ablest teachers can be identified. If there is no satisfactory consistency, that is another story, but at least the effect would be good local propaganda for calling attention to the importance of teaching." (Wolfe, 1964).

In order, therefore, "As a start...," to assist in the objective collection of data related to teaching and instructional effectiveness, a student opinion research program was undertaken by the Senior author in 1961 at The Pennsylvania State University. This first attempt concentrated on the reliable collection of student opinions relative to the instructional environment in which the student operates. Obviously, these data contribute only a part of the available information, and should be considered thus.

Various methods for the collection of student opinion were considered in the development of the final questionnaire form. Guthrie (1954), used a rating system wherein the student responds

to five objective type items, and two essay questions (i.e. what has the instructor done especially well, or what might be done to improve his teaching in this course). The University of Nebraska uses a ranking form on which the student lists 8 other instructors he has had, and compares his current instructor, in rank order, with these eight.

Remmers, in the development of the Purdue Rating Scale for Instruction, used 26 items responded to on the basis of extremely poor, below average, average, above average and excellent Likert scale, or on a sematic differential type response scale (10 positions wide).

A review of the various procedures and forms led first to the designation of the parameters and limitations under which such a rating procedure must conform. The following criteria were developed:

1. It must be objectively scorable, to insure rapid processing and equal treatment per instructor or course.
2. It must gather opinions on those areas of the instructional process which differentiate instructors and programs and to which students develop differential opinions.
3. It must be able to be administered by the instructor himself in regular class time or examination time, so that proctors or test administrators would not be necessary.
4. It must reflect opinions developed within the instruct-

ional situation rather than opinions developed prior to exposure to the course, or due to course content, time of day, required or elective, level of course, etc.

5. It must be reliable (above .90)
6. It must be able to identify "true" student responses, and differentiate or separate out those student responses which seem irresponsible, immature or careless.
7. It must be as diagnostic as possible, so that the results can be adequately interpreted by the instructor, and serve to help him understand the nature and effectiveness of the instructional communication process.
8. It must be confidential; i.e. the results must retain the ananymity of the instructor who is being evaluated, and the anonymity of the student.
9. It must cover those areas of the instructional process which validly relate to learning.
10. It must be long enough to insure reliability, diagnostic capability, and validity.
11. It must differentiate between and among instructors.

In the comparison of existing student opinion forms against the criteria thus established, little data was available. A collection of items and questions was made, in order to determine the areas which seemed to be represented on already existing forms. It was evident that many items in use reflected specific points of view

about a teaching/learning environment.

For example:

The University of Iowa. Survey of Student Opinion of Teaching

10. Personal interest in students and sensitivity to student problems.

Agronomy Department, University of Illinois. Course and Teacher Rating Form (mimeographed)

27. Use of English Language

29. Eye contact (looks directly at class)

22. Use of visual aids (including blackboard).

General Engineering Department, University of Illinois. Instructor Rating Form.

5. Personal Appearance

Always well groomed, usually well groomed, careless about appearance, untidy in appearance, extremely untidy in appearance.

University of Minnesota, Survey of Student Reactions to Courses and Instruction, 1961.

3. What interest would you have in taking other courses in this general area of study?

Remmers, H. H. and Elliott, D. N. The Purdue Rating Scale for Instruction. 1950, The Purdue Research Foundation, Purdue University, Lafayette, Indiana.

4. Liberal and progressive attitude

9. Personal Appearance

19. Freedom allowed students in the selection of materials to be studied.

Reference to existing forms was eliminated as a methodology, and instead, concepts related to student learning were sought. A group of students was asked to make statements describing good and poor teaching (Speech classes, The Pennsylvania State University). Next, a committee of faculty met to independently develop concepts of effective teaching. These student and faculty concepts were re-written to a common format, and grouped by area. Duplicates were eliminated. Some 500 or so statements were the result. These statements were produced in groups of 100 items each, and pilot groups of students were administered one of the forms. The items were analyzed, and reduced in number by eliminating items which were highly skewed (very high agreement or disagreement), or which were not loaded on any identifiable factor (principal components, varimax rotation). The result was a form containing 150 items. These were pre-tested on several small groups, refined, and a second form developed.

During the academic year 1961-62 the second form was administered to 1,319 students in Accounting, Anthropology, Army, History, Mathematics, Music, Psychology and Zoology. At Pennsylvania State the response positions for this second form were "final-choice" responses, since the middle or neutral position on the first form attracted most of the student responses. Secondly many students seemed to respond to the first questionnaire with either a response

set, or sheer boredom, by selecting only the middle position.

The reliability of the second questionnaire was computed with an analysis of variance method (Hoyt, 1941). There were 6 items dealing with the laboratory, 35 course content items and 37 instructional method items, obtaining reliabilities of .64, .90 and .89. The overall reliability for the total questionnaire was .93.

The correlations between the total scores on the three parts of the test were as follows: laboratory and content, r=.343; laboratory and method, r =.267; content and method, r = .636. It would appear from these correlations that the various parts of the questionnaire were measuring different aspects of the course.

The total score for each part of the questionnaire was based upon a score of 4 if the student strongly agreed with a positive statement or strongly disagreed with a negative item. If he only agreed with a positive statement or disagreed with a negative statement, the score for the item would be a 3. The results of the use of this scoring system on the 482 questionnaires for a course in Zoology 25 are presented in Table 3. The maximum number of points for an item was 4; the maximum number of points for any part of the questionnaire was four times the number of items in the part. The heading "% Maximum" indicates the proportion of student agreement with the maximum possible score.

Table 3

Means, Standard Deviations, Maximum Number of Points, Standard Error of Measurement and Per Cent Agreement With Maximum Score for 483 Students in Zoology 25

Questionnaire Part	Mean	St. Dev.	S. E. Meas.	Max # Pts.	% Max.
Laboratory (6 items)	19.6	2.2	1.88	34	.82
Content (34 items)	101.6	10.4	5.01	136	.75
Method (37 items)	101.8	11.2	5.57	148	.69
Total (77 items)	223.0	20.4	7.77	308	.72

A correlational analysis was made of the relationship between the scores on the three parts of the questionnaire and the students' term, sex, expected grade and reason for taking the course (required or elective). Table 4 indicates that the scores do not correlate with term, sex or reason for taking the course; the scores on the content and method sections are significantly correlated with expected grade, p .01, N = 483, but expected grade accounts for only 7 per cent of the variance in the content items and 2 per cent of the variance in the method items.

Table 4

Product-Moment Correlations Between Total Scores for Parts and Biographical Information N = 483

	Term	Sex	Expected Grade	Required-Elective
Laboratory	-.006	.083	.025	-.059
Content Items	-.044	.007	.257*	-.090
Method Items	-.077	.007	.149*	-.032

*p. .01

This means that only a very small portion of the variance in the

students' responses is related to the grade which they expect to receive in the course.

The correlation between the response to any particular item and the total score for that part of the attitude questionnaire can be interpreted as the relationship between what the item measures and what the total score represents. If the correlations are all very high, the items are all measuring nearly the same thing; if the correlations are very low, the items are all measuring something different. It is usually desirable to have items whose correlations with the total score fall somewhere between these extremes. Table 5 shows the average item-total score correlations (r) for the items used in the questionnaire. Only six of the 77 items had item-total score correlations below .25. The table also indicates the average inter-item correlations.

Table 5

Average Item-Total Score Correlations and Inter-Item Correlations for the Laboratory, Content and Method Items N = 483

Laboratory	.598	.23
Content Items	.498	.20
Method Items	.435	.17

A factor analysis of the item correlations produced ten factors which accounted for 47.8 per cent of the variance in

the responses. The factors were tentatively identified as:

1. Content - both general and specific items
2. Method
3. Ease - Pace
4. Additional Materials
5. Television
6. Interest - Attention
7. Laboratory
8. Organization of Material
9. Student Participation
10. Tests

With this data in hand, an attempt was made to reduce the number of items which appeared on a first "general" factor, and increase the number of items on the second order factors. Those items which factored singly were eliminated. Forty-two items resulted from this analysis, and eight of the single type items were retained as "specific items" for which faculty interest was high (tests, textbook, homework, readings, etc.). Half of the items were then made negative, and matched to corresponding, similarly loaded items on the same factor. This enabled the construction of a "lie" or "fallibility" score. (Those items with highest intercorrelations were paired and one made negative.)

The items on course content were separated into two categories: those which seem to reflect a general attitude toward the course and those which specifically refer to content. The

items on method of instruction were made more general in order that they might apply to any course. A number of items were added which had specific reference to the instructor; the items which reflected interest and attention were retained as well as items referring to organization of the course, homework, tests, pace of the course, student participation and outside reading.

A general purpose questionnaire resulted which serves two functions:

1. Comparisons can be made between students' perceptions of a particular course with norms established from other courses throughout the university, and

2. comparisons can be made between different aspects of one course (i.e., content, method, instructor, etc.). The questionnaire's main function is as a diagnostic device to identify what the students believe to be the more outstanding characteristics of the courses which are offered, and instructor capability in presenting that course.

Reliability:

There are various methods of estimating reliability on a measuring instrument of this type. The split half method was performed on the negative vs. positive items on the sample of 297 questionnaires, yeilding a correlation of .849, which corrected for length (Spearman-Brown) = .92. A second split-half reliability was computed with half the negative and half the positive items in each group thus 25 items in each half. The result was

an obtained correlation of .865, which corrected for length = .93.

A mean reliability of 96.1 was obtained (split-half) on 379 class sections. Kuder-Richardson reliabilities were computed on several samples. The KR formulas and underlying concepts provide in this instance an underestimate of the reliability of the questionnaire. The KR 14 assumes that a single common factor is being measured, that all inter item correlations are equal, and that the scoring formula is the number of right responses. KR 20 assumes an additional postulate, that item variances are equal; and KR 21 assumes, in addition, that item difficulties are equal. Secondly, weighted responses are not considered. Considering, therefore, that the most positive response is a "right" answer, and all others wrong, the following results were obtained:

	Sample	
Economics 108	587	KR 14 = .940; KR 20 = .939; KR 21 = .932
Hygine 104	357	KR 14 = .935; KR 20 = .934; KR 21 = .928

Style of Items:

The questionnaire items may seem to appear very brief and stilted. It was found, however, that the shorter and more definite items factored, while items which included more than one concept or element, did not so factor. It was difficult to isolate items which could be definitely assigned to the measurement of specific elements in the instructional program; i.e. content vs. method. It was essential, therefore, to restrict the item content to specific and discrete elements, if the idea of factor

scores was to be retained. Since the object of the questionnaire was to differentially gather student opinion on those variables which they recognize as important elements in the teaching process, factorial scores seemed a most appropriate method for the construction of the questionnaire.

Type of Instrument:

The problem of gathering data on the efficiency of instruction suffers through the lack of objectivity. Objective measures of teaching are not available, so one must resort to systems which are to some degree subjective. One set of data which are available is the opinions of students. It is to be recognized that one essential characteristic of such ratings and rating scales is that they must be reliable if they are to be used in an evaluation program. In general, the more ratings obtained, the more reliable will be the results; or, the more items on a rating scale, the greater is the potentiality for high degrees of reliability. This instrument was designed to achieve reliability coefficients (of the internal homogeniety type) above .90. Secondly, it was constructed in such a way that individual student responses can be evaluated as to their reliability.

Negative and positive questionnaire items:

Each positive item in the 5 major subscore factors is matched by a negative item. For example;

49. The course was quite useful

29. One of my poorest courses

or

22. Held my attention throughout the course

46. It was quite boring

Such a method of developing questionnaire items, recognizable to the student examinee was found to increase reliability. It evidently exerts a pressure on the student to attend more to his answers, prevents random answering patterns, and improves student interest and attention during the questionnaire administration.

On a sample of 297 student questionnaire responses, the standard deviation on the 24 positive items was 12.21, and on the 26 negative items the standard deviation was 12.15, indicating a very common and homogenous answering pattern on both types of items. The Mean positive item score was 2.719, and 2.807 was obtained as the mean negative item score. The hypothetical item mean is 2.5, so that it can be seen that a slight overall positive response set is obtained.

There are two basic demensions necessary in the development of a teacher or course rating scale: (1) the elements that the students respond to (i.e. the items) are know to differentiate among teachers; and (2) norms are developed of a sufficient number and demension to adequately compensate for extraneous, but correlated variables affecting the ratings obtained, and provide useful interpretable comparisons. The Illinois Course Evaluation Questionnaire was developed with these two essential characteris-

tics in mind -- items were selected which yielded the widest obtained variance among instructors, and norms have been collected from over 100,000 cases. Norm tables have been built wherever extraneous factors have been found. Such differences were discovered in course content (mathmatics vs. English), the level of the course (Freshman Rhetoric vs. Advanced Conversational French), rank of the instructor (Professor vs. graduate teaching assistant), and college (Liberal Arts vs. Engineering). Norm tables not compensating for these differences would have to assume, from CEQ data, that mathematics is actually taught better than English, that freshmen level courses are taught poorer than graduate courses, professors teach better than assistant professors. Other factors investigated, but showing no significant difference included size of class (N from 6 to 175), sex of instructor, whether the course was required or elective, or at what hour the course was taught. Other sources of variance now being investigated include grade earned in the course (correlations of approximately .30 are regularly obtained with student report of expected grade), the "mix" of male/female students in the class vs. the sex of the instructor (female students tend to rate female instructors lower than male instructors), and the degree to which individual students may carry a halo evaluation tendency -- a personal response set -- which marks them as particularly critical or praising.